(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号
特開2003－248686
（P2003－248686A）
(43)公開日　平成15年９月５日(2003.9.5)

(51)Int.Cl.⁷		識別記号	FI			テーマコード（参考）
Ｇ０６Ｆ	17/30	２１０	Ｇ０６Ｆ	17/30	２１０Ｄ	５Ｂ００９
		１７０			１７０Ａ	５Ｂ０７５
	17/21	５７０		17/21	５７０Ｒ	

審査請求　未請求　請求項の数11　ＯＬ　（全 11 頁）

(21)出願番号　特願2002－45516(P2002－45516)

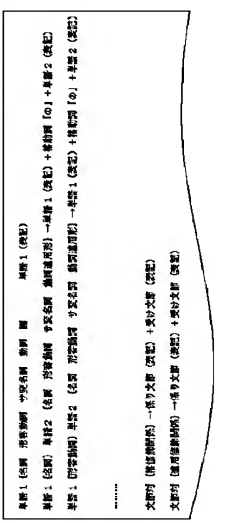
(22)出願日　平成14年２月22日(2002.2.22)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　佐藤　奈穂子
東京都大田区中馬込１丁目３番６号　株式会社リコー内

Ｆターム(参考)　5B009 SA12 SA14 VA02
5B075 ND03 NR12 UU06

(54)【発明の名称】　文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体

(57)【要約】
【課題】本発明は文書群の言語属性を言語解析して文書群の内容を個々に読むことなくその内容を示すラベルを自動的に取得する文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体を提供する。
【解決手段】文書群ラベル生成装置１は、テキスト文書群記憶部２に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群の言語属性を言語解析部３で解析し、解析された言語属性情報を計量して、計量結果に基づいてテキスト文書群に特徴的な言語現象を抽出し、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶部６のラベル生成規則との照合処理を行ってテキスト文書群に対してテキスト文書群の内容を示すラベルをラベル生成部４で生成している。したがって、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得する。



1

【特許請求の範囲】

【請求項１】文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成装置において、前記文書蓄積手段のテキスト文書の言語属性を解析する言語解析手段と、当該言語解析手段の解析した言語属性情報を計量する計量手段と、当該計量結果に基づいて前記テキスト文書群に特徴的な言語現象を抽出する特徴抽出手段と、ラベル生成規則を記憶するラベル生成規則辞書記憶手段と、前記特徴抽出手段の抽出した前記特徴的な言語現象と前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行って前記テキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するラベル生成手段と、を備えていることを特徴とする文書群ラベル生成装置。

【請求項２】前記計量手段は、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量することを特徴とする請求項１記載の文書群ラベル生成装置。

【請求項３】前記特徴抽出手段は、前記計量手段で計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出することを特徴とする請求項１記載の文書群ラベル生成装置。

【請求項４】前記ラベル生成規則辞書記憶手段は、前記ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であることを特徴とする請求項１記載の文書群ラベル生成装置。

【請求項５】前記ラベル生成手段は、前記特徴抽出手段で抽出された対象文書群の特徴的な言語現象を正規化して、前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうことを特徴とする請求項１記載の文書群ラベル生成装置。

【請求項６】文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法において、前記文書蓄積手段のテキスト文書の言語属性を解析する言語解析処理ステップと、当該言語解析処理ステップで解析した言語属性情報を計量する計量処理ステップと、当該計量結果に基づいて前記テキスト文書群に特徴的な言語現象を抽出する特徴抽出処理ステップと、前記特徴抽出処理ステップで抽出した前記特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行って前記テキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するラベル生成処理ステップと、の各ステップ処理を行うことを特徴とする文書群ラベル生成方法。

2

【請求項７】前記文書群ラベル生成方法は、前記計量処理ステップで、少なくとも言語属性情報として各文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量することを特徴とする請求項６記載の文書群ラベル生成方法。

【請求項８】前記文書群ラベル生成方法は、前記特徴抽出処理ステップで、前記計量処理ステップで計量された言語属性情報のうち、頻度の高い情報を、該当文書群における特徴的な言語現象として抽出することを特徴とする請求項６記載の文書群ラベル生成方法。

【請求項９】前記文書群ラベル生成方法は、前記ラベル生成規則辞書記憶手段が、前記ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であることを特徴とする請求項６記載の文書群ラベル生成方法。

【請求項１０】前記文書群ラベル生成方法は、前記ラベル生成処理ステップで、前記特徴抽出処理ステップで抽出された対象文書群の特徴的な言語現象を正規化して、前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうことを特徴とする請求項６記載の文書群ラベル生成方法。

【請求項１１】文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法のプログラムを記録する記録媒体であって、前記請求項６から請求項１０のいずれかに記載の文書群ラベル生成方法のプログラム及びデータを記録することを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体に関し、詳細には、テキスト文書の言語属性を言語解析して、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得する文書群ラベル生成装置、文書群ラベル生成方法及び当該文書群ラベル生成方法とデータを記録した記録媒体に関する。

【０００２】

【従来の技術】近時、情報の電子化が進み、従来では紙文書で保管されていた文書も電子化されるようになってきている。このような文書の電子化に伴って、大量の電子化文書が流通し、収集・蓄積された電子化文書をいかに管理して簡便に再利用するかが重量な問題となってきている。そこでは、ある目的で収集された文書群の自動分類が要望されている。この文書群の自動分類技術は、大量の電子化文書群から類似した文書を自動分類する技術であり、一般的には、各文書に含まれている重要語句

の類似性、出現頻度、出現場所等の共通点に基づいて、関連性の高い文書をグルーピングする技術である。

【０００３】そして、このように分類された文書群を再利用しようとする場合、文書群に対して検索する、一覧する等の行為が生じる。この場合、文書群にラベルが付与されていると、検索を行うことも一覧することも容易である。

【０００４】ところが、一文書には、タイトルや場合によってはキーワードが付与されてラベル付与をするには、現在のところグルーピング後にその文書群の内容を人手で把握して、ラベル付与することになる。

【０００５】そして、テキスト単位群に名前を付与する技術として、従来、テキスト情報群より指定された分析の単位となるテキスト単位群を作成するテキスト情報変換手段と、該作成したテキスト単位群から単語を抽出し、該抽出した単語・テキスト単位間、単語・単語間、テキスト単位・テキスト単位間のうち少なくとも一つの間の距離を計算する距離計算手段と、該計算した距離情報をもとに分析を行う分析手段とを備えたテキスト情報の分析装置が提案されている（特開平１１－３４５２４１号公報参照）。

【０００６】すなわち、この従来技術は、テキスト単位群にユーザが情報の組み合わせや書式を指定して名前を付与している。

【０００７】また、従来、データから特徴的な概念を取り出す、データ分析システムであって、文書データを含むデータからカテゴリ別の概念を抽出する、概念抽出手段と、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する特徴的概念抽出手段を有するデータ分析システムが提案されている（特開２００１－７５９６６号公報参照）。

【０００８】すなわち、この従来技術は、特定の用途向けに予めカテゴリ辞書を用意し、前処理で自動的にデータをラベル付きデータに変換している。

【０００９】

【発明が解決しようとする課題】しかしながら、このような従来技術にあっては、簡単かつ容易に文書を分類して、利用性を向上させる上で改良の必要があった。

【００１０】すなわち、上記特開平１１－３４５２４１号公報記載の従来技術にあっては、テキスト文書群が多数ある場合、ユーザが各々のテキスト文書群に対して指定を行う必要があり、ユーザに非常な労力を強いることとなり、改良の必要があった。

【００１１】また、特開２００１－７５９６６号公報記載の従来技術にあっては、特定の用途向けにカテゴリ辞書を用意する必要があり、その辞書の構築に高額な費用を要するだけでなく、別の用途に再利用しにくいという問題があった。

【００１２】ところで、ラベル付与の自動化を目的とする場合、グルーピング時の検索条件式をそのままラベル化するという方法が一般的であるが、一般に、検索条件式は、単語やキーワードのアンド、オアによる組み合わせが多いため、検索式をそのままを表示しても、ユーザがその文書群の内容を把握するのは困難であり、さらに、表示時の一覧性という観点で適切ではない。

【００１３】一方、重要文抽出技術や要約技術が文書内容の概要を知るための一つの手段として利用することができるが、文書群の特徴を知るために、各々の文書の重要文や要約を読む必要があり、ユーザにとっては非常な労力を強いられることとなり、改良の必要があった。

【００１４】そこで、請求項１記載の発明は、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、文書蓄積手段のテキスト文書の言語属性を言語解析手段で解析し、当該解析された言語属性情報を計量手段で計量して、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を特徴抽出手段で抽出し、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルをラベル生成手段で生成することにより、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得し、利用性の良好な文書群ラベル生成装置を提供することを目的としている。

【００１５】請求項２記載の発明は、計量手段が、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量することにより、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得し、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成装置を提供することを目的としている。

【００１６】請求項３記載の発明は、特徴抽出手段が、計量手段で計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出することにより、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行い、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成装置を提供することを目的としている。

【００１７】請求項４記載の発明は、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であることにより、さまざまな言語表現を吸収すると

5

ともに、ユーザ所望のラベル形式を設定し、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成装置を提供することを目的としている。

【００１８】請求項５記載の発明は、ラベル生成手段が、特徴抽出手段で抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうことにより、ラベル生成規則辞書をより一層有効に利用可能とし、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成装置を提供することを目的としている。

【００１９】請求項６記載の発明は、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、言語解析処理ステップで、文書蓄積手段のテキスト文書の言語属性を解析し、計量処理ステップで、当該解析された言語属性情報を計量して、特徴抽出処理ステップで、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を抽出し、ラベル生成処理ステップで、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成することにより、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得し、利用性の良好な文書群ラベル生成方法を提供することを目的としている。

【００２０】請求項７記載の発明は、計量処理ステップで、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量することにより、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得し、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成方法を提供することを目的としている。

【００２１】請求項８記載の発明は、特徴抽出処理ステップで、計量処理ステップによって計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出することにより、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行い、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成方法を提供することを目的としている。

【００２２】請求項９記載の発明は、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則と

6

が複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であることにより、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定し、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成方法を提供することを目的としている。

【００２３】請求項１０記載の発明は、ラベル生成ステップで、特徴抽出処理ステップにより抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうことにより、ラベル生成規則辞書をより一層有効に利用可能とし、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性の良好な文書群ラベル生成方法を提供することを目的としている。

【００２４】請求項１１記載の発明は、記録媒体に、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法のプログラムであって、請求項６から請求項１０のいずれかに記載の文書群ラベル生成方法のプログラム及びデータを記録することにより、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得し、利用性を向上させることのできる文書群ラベル生成装置及び文書群ラベル生成方法を実現する記録媒体を提供することを目的としている。

【００２５】

【課題を解決するための手段】請求項１記載の発明の文書群ラベル生成装置は、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成装置において、前記文書蓄積手段のテキスト文書の言語属性を解析する言語解析手段と、当該言語解析手段の解析した言語属性情報を計量する計量手段と、当該計量結果に基づいて前記テキスト文書群に特徴的な言語現象を抽出する特徴抽出手段と、ラベル生成規則を記憶するラベル生成規則辞書記憶手段と、前記特徴抽出手段の抽出した前記特徴的な言語現象と前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行って前記テキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するラベル生成手段と、を備えていることにより、上記目的を達成している。

【００２６】上記構成によれば、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、文書蓄積手段のテキスト文書の言語属性を言語解析手段で解析し、当該解析された言語属性情報を計量手段で計量して、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を特徴抽出手段で抽出し、抽出された特徴的な言語現

象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルをラベル生成手段で生成するので、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることができる。

【００２７】この場合、例えば、請求項２に記載するように、前記計量手段は、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量するものであってもよい。

【００２８】上記構成によれば、計量手段が、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量するので、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００２９】また、例えば、請求項３に記載するように、前記特徴抽出手段は、前記計量手段で計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出するものであってもよい。

【００３０】上記構成によれば、特徴抽出手段が、計量手段で計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出するので、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行うことができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００３１】さらに、例えば、請求項４に記載するように、前記ラベル生成規則辞書記憶手段は、前記ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能なものであってもよい。

【００３２】上記構成によれば、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であるので、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００３３】また、例えば、請求項５に記載するように、前記ラベル生成手段は、前記特徴抽出手段で抽出された対象文書群の特徴的な言語現象を正規化して、前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうものであってもよい。

【００３４】上記構成によれば、ラベル生成手段が、特徴抽出手段で抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうので、ラベル生成規則辞書をより一層有効に利用可能とすることができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００３５】請求項６記載の発明の文書群ラベル生成方法は、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法において、前記文書蓄積手段のテキスト文書の言語属性を解析する言語解析処理ステップと、当該言語解析処理ステップで解析した言語属性情報を計量する計量処理ステップと、当該計量結果に基づいて前記テキスト文書群に特徴的な言語現象を抽出する特徴抽出処理ステップと、前記特徴抽出処理ステップで抽出した前記特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行って前記テキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するラベル生成処理ステップと、の各ステップ処理を行うことにより、上記目的を達成している。

【００３６】上記構成によれば、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、言語解析処理ステップで、文書蓄積手段のテキスト文書の言語属性を解析し、計量処理ステップで、当該解析された言語属性情報を計量して、特徴抽出処理ステップで、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を抽出し、ラベル生成処理ステップで、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するので、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることができる。

【００３７】この場合、例えば、請求項７に記載するように、前記文書群ラベル生成方法は、前記計量処理ステップで、少なくとも言語属性情報として各文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量してもよい。

【００３８】上記構成によれば、計量処理ステップで、

少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量するので、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００３９】また、例えば、請求項８に記載するように、前記文書群ラベル生成方法は、前記特徴抽出処理ステップで、前記計量処理ステップで計量された言語属性情報のうち、頻度の高い情報を、該当文書群における特徴的な言語現象として抽出してもよい。

【００４０】上記構成によれば、特徴抽出処理ステップで、計量処理ステップによって計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出するので、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行うことができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００４１】さらに、例えば、請求項９に記載するように、前記文書群ラベル生成方法は、前記ラベル生成規則辞書記憶手段が、前記ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であってもよい。

【００４２】上記構成によれば、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であるので、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００４３】また、例えば、請求項１０に記載するように、前記文書群ラベル生成方法は、前記ラベル生成処理ステップで、前記特徴抽出処理ステップで抽出された対象文書群の特徴的な言語現象を正規化して、前記ラベル生成規則辞書記憶手段の前記ラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なってもよい。

【００４４】上記構成によれば、ラベル生成ステップで、特徴抽出処理ステップにより抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうので、ラベル生成規則辞書をより一層有効に利用可能とすることができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００４５】請求項１１記載の発明の記録媒体は、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法のプログラムを記録する記録媒体であって、前記請求項６から請求項１０のいずれかに記載の文書群ラベル生成方法のプログラム及びデータを記録することにより、上記目的を達成している。

【００４６】上記構成によれば、記録媒体に、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法のプログラムであって、請求項６から請求項１０のいずれかに記載の文書群ラベル生成方法のプログラム及びデータを記録しているので、記録媒体を、コンピュータ等の情報処理装置に読み取らせることで、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることのできる文書群ラベル生成方法を実現する文書群ラベル生成装置を構築することができ、文書群に適切にラベルを生成して、利用性を向上させることができる。

【００４７】

【発明の実施の形態】以下、本発明の好適な実施の形態を添付図面に基づいて詳細に説明する。なお、以下に述べる実施の形態は、本発明の好適な実施の形態であるから、技術的に好ましい種々の限定が付されているが、本発明の範囲は、以下の説明において特に本発明を限定する旨の記載がない限り、これらの態様に限られるものではない。

【００４８】図１～図３は、本発明の文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体の一実施の形態を示す図であり、図１は、本発明の文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体の一実施の形態を適用した文書群ラベル生成装置１のブロック構成図である。

【００４９】図１において、文書群ラベル生成装置１は、テキスト文書群記憶部２、言語解析部３、ラベル生成部４、言語解析用辞書記憶部５及びラベル生成規則辞書記憶部６等を備えている。

【００５０】文書群ラベル生成装置１は、文書分析処理プログラム及び必要なデータを記録するＣＤ－ＲＯＭ（Compact Disc Read Only Memory ）等の記録媒体を、例えば、コンピュータ等に読み取らせて導入することで、構築される。

【００５１】テキスト文書群記憶部（文書群蓄積手段）２は、収集されたテキスト文書のテキスト文書群が登録され、登録されたテキスト文書群がラベル生成の対象と

なる。

【００５２】言語解析用辞書記憶部５は、言語解析部３による言語解析に必要な各種言語解析情報を記憶する。

【００５３】言語解析部（言語解析手段、計量手段、特徴抽出手段）３は、言語解析用辞書記憶部５の記憶する言語解析用辞書に基づいて、テキスト文書群記憶部２によりテキスト文書群記憶部２に登録された各テキスト文書を言語解析用辞書記憶部５の言語解析情報に基づいて言語解析する言語解析処理、言語解析結果における言語属性情報を計量する計量処理及び計量処理の計量結果に基づいてテキスト文書群の特徴的な言語現象を抽出する特徴抽出処理等の各ステップ処理を実行して、処理結果をラベル生成部４に出力する。

【００５４】ラベル生成規則辞書記憶部（ラベル生成規則辞書記憶手段）６は、ラベル生成規則を保持し、例えば、図２に示すようなもので、正規化された言語現象と当該言語現象をラベル化するための規則が複数パターンにわたって記録されている。また、ラベル生成規則辞書記憶部６には、ユーザが規則を新規に適宜追加登録可能である。

【００５５】ラベル生成部（ラベル生成手段）４は、言語解析部３の特徴抽出処理で抽出された特徴的な言語現象とラベル生成規則辞書記憶部６に登録されているラベル生成規則辞書のラベル生成規則とのマッチング処理を行って、対象文書群に対して、そのマッチング内容を示すラベルを生成する。ラベル生成部４は、具体的には、例えば、言語解析部３の特徴抽出処理で抽出された特徴的な言語現象を正規化し、ラベル生成規則辞書記憶部６を検索して、一致した言語現象に対してラベル化を行って、ラベルを生成する。

【００５６】次に、本実施の形態の作用を説明する。文書群ラベル生成装置１は、文書分析処理プログラム及び必要なデータを記録するＣＤ－ＲＯＭ等の記録媒体を、例えば、コンピュータ等に読み取らせて導入することで、構築され、電子化されたテキスト文書群を言語解析して言語属性情報を計量し、テキスト文書群の特徴的な言語現象を抽出して、抽出された特徴的な言語現象とラベル生成規則とのマッチング処理を行って、対象文書群に対して、そのマッチング内容を示すラベルを生成する。

【００５７】すなわち、文書群ラベル生成装置１は、分析対象のテキスト文書群が入力されると、当該テキスト文書群をテキスト文書群記憶部２に登録する。

【００５８】そして、文書群ラベル生成装置１は、図３に示すように、言語解析部３が、言語解析用辞書記憶部５の記憶する言語解析用辞書に基づいて、テキスト文書群記憶部２に登録された各テキスト文書を言語解析、例えば、形態素解析を行なって、その品詞等の属性情報を得たり、係り受け解析を行なって、係り受けの関係のある文節対を得たり、書き手の意図を推定できる語句を得たりする言語解析を行う言語解析処理を行い（ステップＳ１０１）、言語解析の解析結果における言語属性情報を計量する計量処理を行う（ステップＳ１０２）。

【００５９】次に、言語解析部３が、計量処理の計量結果に基づいてテキスト文書群の特徴的な言語現象を抽出する特徴抽出処理を実行して、処理結果をラベル生成部４に出力する（ステップＳ１０３）。

【００６０】次に、ラベル生成部４が、言語解析部３の特徴抽出処理で抽出された特徴的な言語現象に基づいて、ラベル生成規則辞書記憶部６に登録されているラベル生成規則の辞書引きを行う辞書引き処理を行い（ステップＳ１０４）、特徴抽出処理で抽出された特徴的な言語現象とマッチングするラベルを生成するラベル生成処理を行う（ステップＳ１０５）。

【００６１】そして、いま、例えば、ある海のスポーツについて意見を収集・蓄積したテキストデータがあり、集めた意見を内容別に分類し、それぞれのグループに適したラベルを付与して整理する場合、まず、最初に、全てのテキストデータを内容別にグルーピングする。テキストデータを内容別にグループ分けするには、既存の文書検索技術、文書分類技術、クラスタリング技術等を用いて行うことができる。このグルーピングの結果、以下のような文書群Ａ～Ｄが得られたとものとする。

【００６２】〈文書群Ａ〉
・どこでもできるような気がしない。とても楽しそうだが面倒くさそう。
・仲間と楽しく遊びたい。
・楽しそうだけど自分にはちょっと向いていない気がする。
・きっかけがないという感じです。でもやりたいです。たのしそう。

【００６３】〈文書群Ｂ〉
・お金がかかる。
・もっと余暇と、お金があればもっと楽しめると思うが、やりたくてもできないな。
・おもしろそーでやりたいけどお金がかかりそう。

【００６４】〈文書群Ｃ〉
・夏しかできない気がする。ボツンとあってさみしい。
・夏にうってつけの遊び。楽しいぞ。
・夏ならではのたのしい遊び。

【００６５】〈文書群Ｄ〉
・安く手軽にできるならやってみたい。
・もう少し手軽にできないものなのかな。日本だと何かと制限とかうるさそうなので。
・手軽に出来ない。でもぜひイルカと一緒に泳ぎたい。
・ ジェットスキーなどもう少し手軽にできるようになるといい。

【００６６】文書群ラベル生成装置１は、これらのテキスト文書群それぞれに対して、言語解析部３で、言語解析、例えば、形態素解析を行なって、その品詞等の属性

情報を得たり、係り受け解析を行なって、係り受けの関係のある文節対を得たり、書き手の意図を推定できる語句を得たりする言語解析を行う。これら言語解析は、既存のさまざまな手法で実現可能である。

【００６７】さらに、言語解析部３は、テキスト文書群毎に、これらの出現頻度を計量し、頻出語句について、一定のフィルタリングを行ない、その文書群に特徴的な語句を抽出する。この計量処理で計量対象となる単位は、単語、文節、または、係り受け対等のように任意に設定することができる。また、特徴語句のフィルタリングは、情報検索技術で用いられている品詞限定や不要語除去等の手法を用いて実現することができる。

【００６８】そして、言語解析部３で、上記例のテキスト文書群について、各文書群に出現する語句の計量とその頻出語句を品詞によってフィルタリングを行なったところ、特徴語句として、以下の情報（特徴的な言語現象）が抽出された。

【００６９】文書群Ａ：「楽しい（形容詞）」
文書群Ｂ：「お金（名詞）が」→「かかる（動詞）」
文書群Ｃ：「夏（名詞）」「遊び（名詞）」
文書群Ｄ：「手軽（形容動詞）に」→「できる（助動詞）＋ない（助動詞）」

次に、抽出された語句（言語現象）を、ラベル生成規則辞書記憶部６に登録されているラベル生成規則辞書で検索可能な形式に変換する。この場合、ラベル生成規則辞書として、図２に示したようなラベル生成規則辞書を用いるとすると、抽出された語句は、以下のように変換される。これらは、言語解析の結果得られた語句の属性情報の並び替えや正規化で行われる。

【００７０】
文書群Ａ：単語１｛形容詞｝　単語１表記（楽しい）
文書群Ｂ：文節対｛格修飾関係｝　係り文節表記（お金が）→受け文節表記（かかる）
文書群Ｃ：単語１｛名詞｝　単語２｛名詞｝　単語１表記（夏）単語２表記（遊び）
文書群Ｄ：文節対｛連用修飾関係｝　係り文節表記（手軽に）→受け文節表記（できない）

そして、ラベル生成部４が、上記形式で、図２に示したラベル生成規則辞書を適用し、一致した言語現象に対してラベル化を行ったところ、以下のラベルを得ることができた。

【００７１】文書群Ａ：ラベル（楽しい）
文書群Ｂ：ラベル（お金がかかる）
文書群Ｃ：ラベル（夏の遊び）
文書群Ｄ：ラベル（手軽にできない）

すなわち、上記例では、ある海のスポーツについての意見を内容別に分類すると、「楽しい」「お金がかかる」「夏の遊び」「手軽にできない」と整理することができ、各文書群を再利用可能なラベル付きデータ群として保存することができる。

【００７２】もし、ユーザが、この例のように、ある海のスポーツの印象についてのアンケートを行なおうとする場合、上記文書群ラベル生成装置１で作成された各文書群のラベルをそのまま選択項目として再利用することができる。

【００７３】このように、本実施の形態の文書群ラベル生成装置１及び文書群ラベル生成方法は、テキスト文書群記憶部２に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、テキスト文書群記憶部２のテキスト文書の言語属性を言語解析部３で解析し、解析された言語属性情報を計量して、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を抽出し、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶部６のラベル生成規則との照合処理を行ってテキスト文書群に対してテキスト文書群の内容を示すラベルをラベル生成部４で生成している。

【００７４】したがって、大量の文書群の内容を個々に読むことなく、収集・蓄積された大量のテキスト文書データをユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、このラベルは、従来のような単語キーワードの域を超えたユーザにとって理解しやすいものである。その結果、利用性を向上させることができる。

【００７５】また、本実施の形態の文書群ラベル生成装置１及び文書群ラベル生成方法は、言語解析部３が、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量している。

【００７６】したがって、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００７７】さらに、本実施の形態の文書群ラベル生成装置１及び文書群ラベル生成方法は、言語解析部３が、計量処理で計量した言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出している。

【００７８】したがって、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行うことができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００７９】また、本実施の形態の文書群ラベル生成装置１及び文書群ラベル生成方法は、ラベル生成規則辞書記憶部６が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象

をラベル化するための規則をユーザが適宜追加登録可能である。

【００８０】したがって、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８１】さらに、本実施の形態の文書群ラベル生成装置１及び文書群ラベル生成方法は、ラベル生成部４が、言語解析部３の特徴抽出処理で抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なっている。

【００８２】したがって、ラベル生成規則辞書をより一層有効に利用可能とすることができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８３】以上、本発明者によってなされた発明を好適な実施の形態に基づき具体的に説明したが、本発明は上記のものに限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【００８４】

【発明の効果】請求項１記載の発明の文書群ラベル生成装置によれば、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、文書蓄積手段のテキスト文書の言語属性を言語解析手段で解析し、当該解析された言語属性情報を計量手段で計量して、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を特徴抽出手段で抽出し、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルをラベル生成手段で生成するので、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることができる。

【００８５】請求項２記載の発明の文書群ラベル生成装置によれば、計量手段が、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量するので、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８６】請求項３記載の発明の文書群ラベル生成装置によれば、特徴抽出手段が、計量手段で計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出するので、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行うことができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８７】請求項４記載の発明の文書群ラベル生成装置によれば、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であるので、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８８】請求項５記載の発明の文書群ラベル生成装置によれば、ラベル生成手段が、特徴抽出手段で抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうので、ラベル生成規則辞書をより一層有効に利用可能とすることができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００８９】請求項６記載の発明の文書群ラベル生成方法によれば、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成するに際して、言語解析処理ステップで、文書蓄積手段のテキスト文書の言語属性を解析し、計量処理ステップで、当該解析された言語属性情報を計量して、特徴抽出処理ステップで、当該計量結果に基づいてテキスト文書群に特徴的な言語現象を抽出し、ラベル生成処理ステップで、抽出された特徴的な言語現象とラベル生成規則を記憶するラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行ってテキスト文書群に対して当該テキスト文書群の内容を示すラベルを生成するので、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることができる。

【００９０】請求項７記載の発明の文書群ラベル生成方法によれば、計量処理ステップで、少なくとも言語属性情報として各テキスト文書毎の単語の出現頻度、出現文節頻度、係り受け関係にある文節対の出現頻度を計量するので、さまざまな言語単位による計量を行なって、文書の特徴をさまざまな言語単位で取得することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００９１】請求項８記載の発明の文書群ラベル生成方法によれば、特徴抽出処理ステップで、計量処理ステップによって計量された言語属性情報のうち、頻度の高い情報を、当該テキスト文書群における特徴的な言語現象として抽出するので、多く存在する言語情報から文書中の高頻度語句を文書群の特徴語句として同定して、容易に特徴語句の絞り込みを行うことができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００９２】請求項９記載の発明の文書群ラベル生成方法によれば、ラベル生成規則辞書記憶手段が、ラベル生成規則辞書として、正規化された言語現象と当該言語現象をラベル化するための規則とが複数パターンにわたって記載されており、当該言語現象をラベル化するための規則をユーザが適宜追加登録可能であるので、さまざまな言語表現を吸収するとともに、ユーザ所望のラベル形式を設定することができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００９３】請求項１０記載の発明の文書群ラベル生成方法によれば、ラベル生成ステップで、特徴抽出処理ステップにより抽出された対象文書群の特徴的な言語現象を正規化して、ラベル生成規則辞書記憶手段のラベル生成規則との照合処理を行い、一致した言語現象に対してラベル化を行なうので、ラベル生成規則辞書をより一層有効に利用可能とすることができ、ユーザにより一層分かりやすい表現による内容を示すラベルを自動的に取得して、より一層利用性を向上させることができる。

【００９４】請求項１１記載の発明の記録媒体によれば、記録媒体に、文書蓄積手段に収集・蓄積された複数のテキスト文書からなる複数のテキスト文書群に対してラベルを生成する文書群ラベル生成方法のプログラムであって、請求項６から請求項１０のいずれかに記載の文書群ラベル生成方法のプログラム及びデータを記録しているので、記録媒体を、コンピュータ等の情報処理装置に読み取らせることで、大量の文書群の内容を個々に読むことなく、ユーザに分かりやすい表現による内容を示すラベルを自動的に取得することができ、利用性を向上させることのできる文書群ラベル生成方法を実現する文書群ラベル生成装置を構築することができ、文書群に適切にラベルを生成して、利用性を向上させることができる。

【図面の簡単な説明】

【図１】本発明の文書群ラベル生成装置、文書群ラベル生成方法及び記録媒体の一実施の形態を適用した文書群ラベル生成装置の要部ブロック構成図。

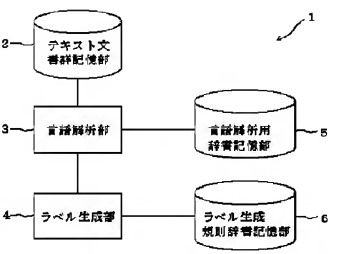
【図２】図１のラベル生成規則辞書記憶部に登録されているラベル生成規則辞書の一例を示す図。

【図３】図１の文書群ラベル生成装置による文書ラベル生成処理を示すフローチャート。

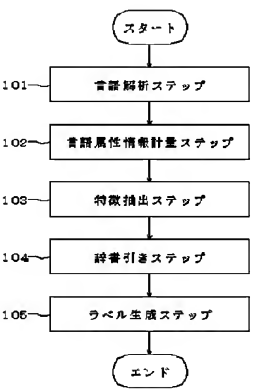
【符号の説明】

１　文書群ラベル生成装置
２　テキスト文書群記憶部
３　言語解析部
４　ラベル生成部
５　言語解析用辞書記憶部
６　ラベル生成規則辞書記憶部

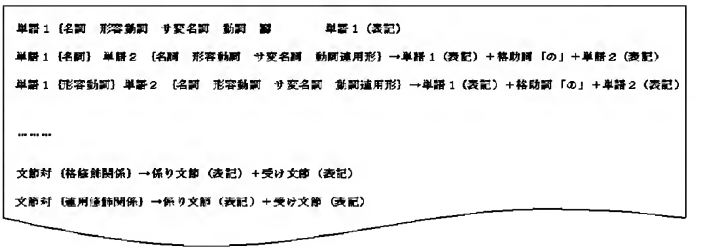
【図１】



【図２】



【図３】


単語１｛名詞　形容動詞　サ変名詞　動詞　語　　単語１（表記）
単語１｛名詞｝単語２｛名詞　形容動詞　サ変名詞　動詞連用形｝→単語１（表記）＋格助詞「の」＋単語２（表記）
単語１｛形容動詞｝単語２｛名詞　形容動詞　サ変名詞　動詞連用形｝→単語１（表記）＋格助詞「の」＋単語２（表記）
……
文節対（格係能関係）→係り文節（表記）＋受け文節（表記）
文節対（連用修飾関係）→係り文節（表記）＋受け文節（表記）

PAT-NO: JP02003248686A
DOCUMENT-IDENTIFIER: JP 2003248686 A
TITLE: DOCUMENT GROUP LABEL CREATION DEVICE AND METHOD, AND RECORDING MEDIUM
PUBN-DATE: September 5, 2003

INVENTOR-INFORMATION:

NAME	COUNTRY
SATO, NAOKO	N/A

ASSIGNEE-INFORMATION:

NAME	COUNTRY
RICOH CO LTD	N/A

APPL-NO: JP2002045516
APPL-DATE: February 22, 2002

INT-CL (IPC): G06F017/30 , G06F017/21

ABSTRACT:

PROBLEM TO BE SOLVED: To provide a document group label creation device and a method linguistically analyzing language attributes of document groups and automatically providing a label showing its contents without individually reading the contents of the document groups and to provide a recording medium.

SOLUTION: This document group label creation device 1 allows a language analysis part 3 to analyze the language attributes of a plurality of text document groups comprising a plurality of text documents collected and stored in a text document group storage part 2, measures the analyzed language attribute information, selects language phenomena characteristic to the text document group based on the measured result, collates the selected characteristic language phenomena with a label creation rule of a label creation rule dictionary storage part 6 storing the label creation rule, and creates the label showing the contents of the text document group for the text document group by a label creation part 4. This constitution thus automatically provides the label showing the content by an expression easily understood by the user without individually reading the contents of the large amount of document groups.